De-identification & Preventing Re-identification

Amy Hawn Nelson, Research Faculty, Director of Training and Technical Assistance, AISP

Sean Cottrell, Operations Director, DISC

Laia Tiderman, Associate Director, DISC

What We Do

- **Convene** and advocate on behalf of communities that are sharing and using cross-sector data for good
- **Connect** to innovations, best practices, and research and funding opportunities that support ethical data sharing
- **Consult** with data sharing collaborations to build the human and technical capacity to share data and improve lives

Why We Do It

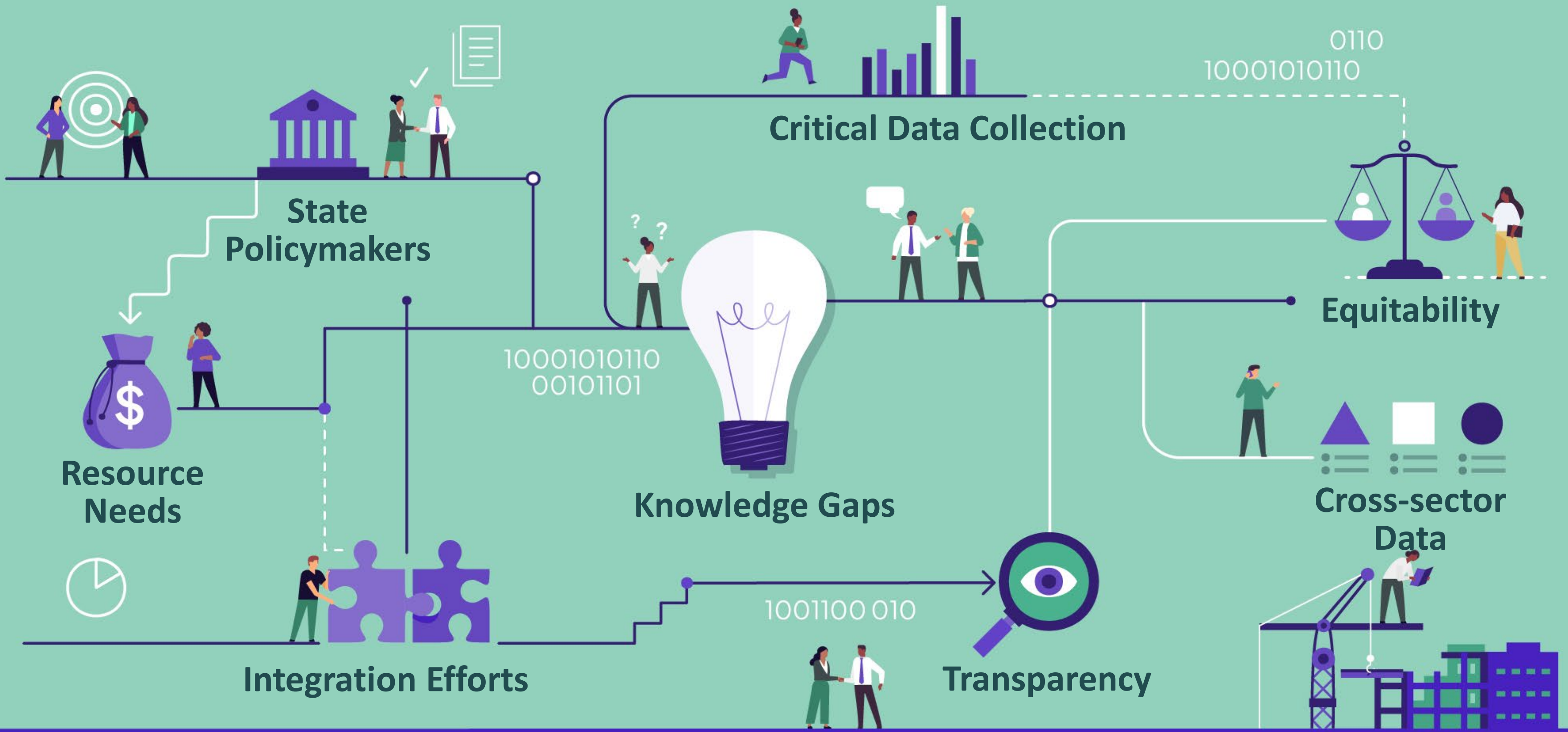
When communities bring together cross-sector data safely and responsibly, policy-makers, practitioners, and schools are better equipped to:

- Understand the complex needs of individuals and families
- Allocate resources where they're needed most to improve services
- Measure long-term and two-generation impacts of policies and programs
- Engage in transparent, shared decision-making about how data should (and should not) be used

www.aisp.upenn.edu

AISP

The Data Integration Support Center (DISC) at WestEd provides expert integrated data system planning and user-centered design, policy, privacy, and legal assistance for public agencies nationwide.

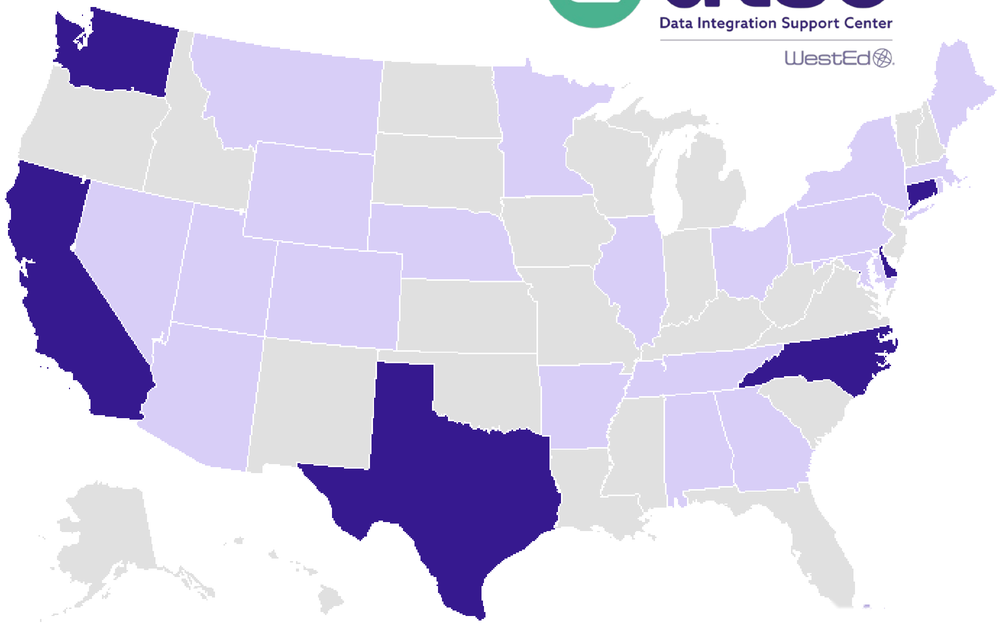


Our roles

We are:	We are not:
Data evangelists	Data holders or intermediaries
Connectors, community builders, thought partners, cheerleaders, and data sharing therapists	A vendor or vendor recommenders
Focused on ethical data use for policy change	Focused on academic research

Our Networks

★ Statewide ◆ Local



● Intensive support ● Moderate support

Our approach

Data sharing is as relational as it is technical.

We don't just need to integrate data;

we need to integrate people.

LEGAL DISCLAIMER

- Not Legal Advice
- Training will only cover **federal law**
- Laws change. This content is based on the law at the time of the workshop
- Consult your general counsel for specific legal questions

Essential Questions

What are the key techniques and methodologies for effectively de-identifying sensitive information to ensure compliance with legal and regulatory standards?

How can lawyers identify and mitigate potential risks of re-identification, and what best practices should be followed to maintain the privacy and confidentiality of client data?

What are the legal and ethical considerations surrounding data de-identification, and how can lawyers navigate these to protect sensitive information while fulfilling their professional responsibilities?

Key Terms

Privacy

Individual autonomy and each person's control over their own information including each person's right to decide when and whether to share personal information, how much information to share, and the circumstances under which that information can be shared

Confidentiality

Management of another individual's personally identifiable information defined as referring to the **obligations of those who receive** personal **information** about an individual to respect the individual's privacy by safeguarding the information

Disclosure

the release or **exposure** of information that was supposed to be confidential

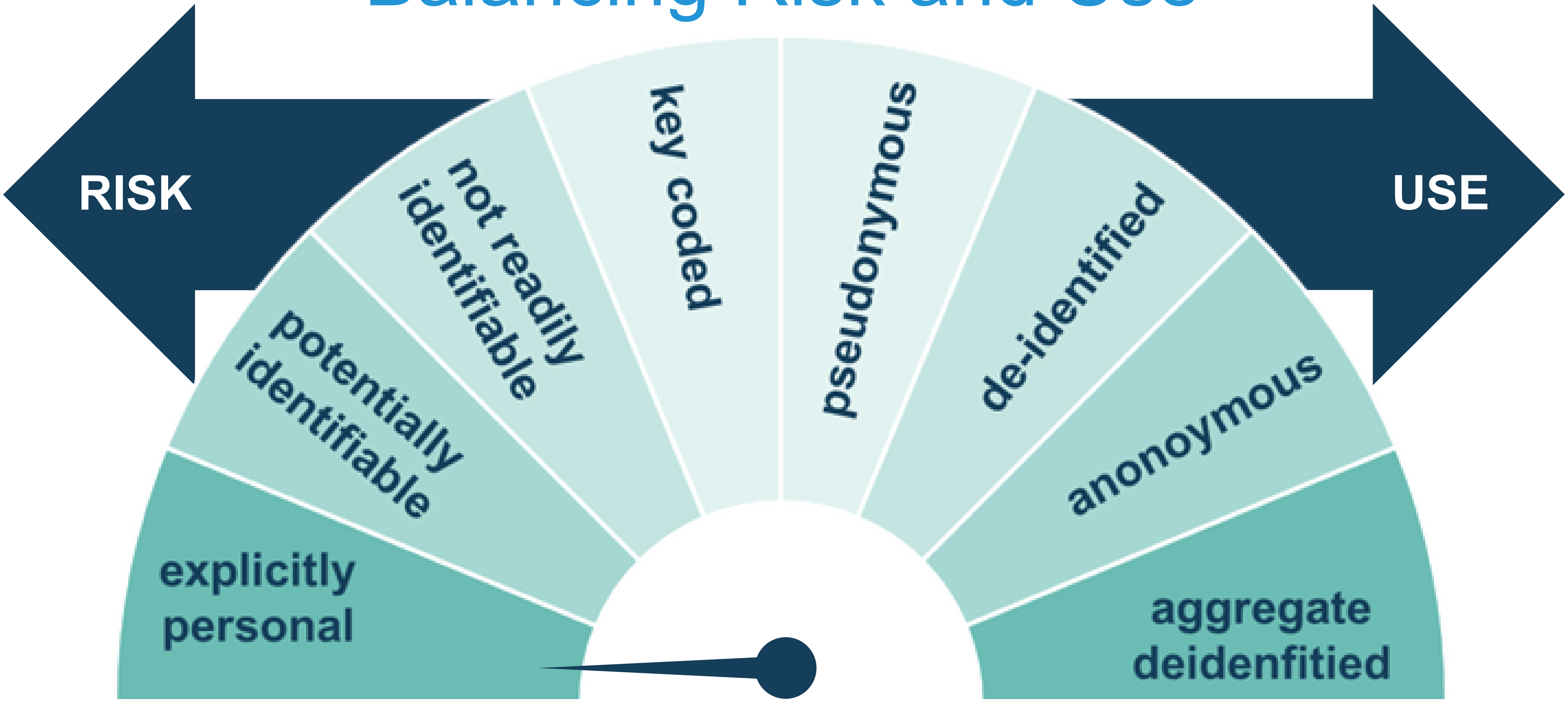
De-identification

refers to the **process** of removing or obscuring any personally identifiable information from a data set, report, or other product in a way that minimizes the risk of unintended disclosure of the identity of individuals and information about them

Re-identification

The matching of de-identified data back to an individual

Balancing Risk and Use



Adapted from [A Visual Guide to Practical De-Identification](), FPF.org

Potential Risks

Re-identification

Risk of re-identification where individuals can be traced back to their data using available or additional information.

Loss of Data Utility

Data losing its usefulness for legitimate analysis, as too many details are stripped away, making it difficult to draw meaningful conclusions.

Data Integrity

Affects to the accuracy or integrity of the data, leading to incorrect analyses or decisions based on flawed information.

Security Vulnerabilities

If proper security measures are not applied post-de-identification, the data might be exposed to unauthorized access or data breaches.

Ethical Concerns

De-identification methods may inadvertently introduce or perpetuate biases of specific groups or communities.

When and why do IDSs need to protect confidentiality?

LEGAL REQUIREMENTS

- Federal, state, and local laws & regulations
- Policies and procedures

ETHICAL OBLIGATIONS

- Role and responsibility as data stewards
- Professional codes of conduct

OPERATIONAL CONSIDERATIONS

- Technical structures to support legal requirements and ethical obligations

Adapted from Federal Committee on Statistical Methodology (2024), Data Protection Toolkit, Legal, Ethical, and Operational Underpinnings of Confidentiality

Legal Standards

Four Questions to consider throughout this work

Is it legal?

Is it ethical?

Is it a good idea?

How do we know?
Who decides?

[Finding a Way Forward: How to create a strong legal framework for data integration, 2022]
Four Questions to Guide Decision-Making for Data Sharing and Integration, 2023, https://ijpds.org/article/view/2159

Balancing Act

- **It is not possible to completely eliminate the risk of disclosure.**
- **Agencies releasing information are responsible for minimizing any such risk while meeting legal standards.**

Federal Privacy Standards

FERPA	• “Reasonable person” standard
HIPAA	• Safe Harbor and Expert Determination
Other Federal Laws	• Higher Education Act • Workforce Innovation and Opportunity Act
State Laws	• State Privacy Laws • State Consumer Protection Laws

Ethical Considerations

Naming some tension in this work

There are significant privacy risks to the reuse and disclosure of individual-level data

There are significant benefits to individuals and communities when we can use individual-level data to improve programs, services, and policies

Risk vs. Benefit Matrix

1: High benefit, low risk

2: High risk, high benefit

3: Low risk, low benefit

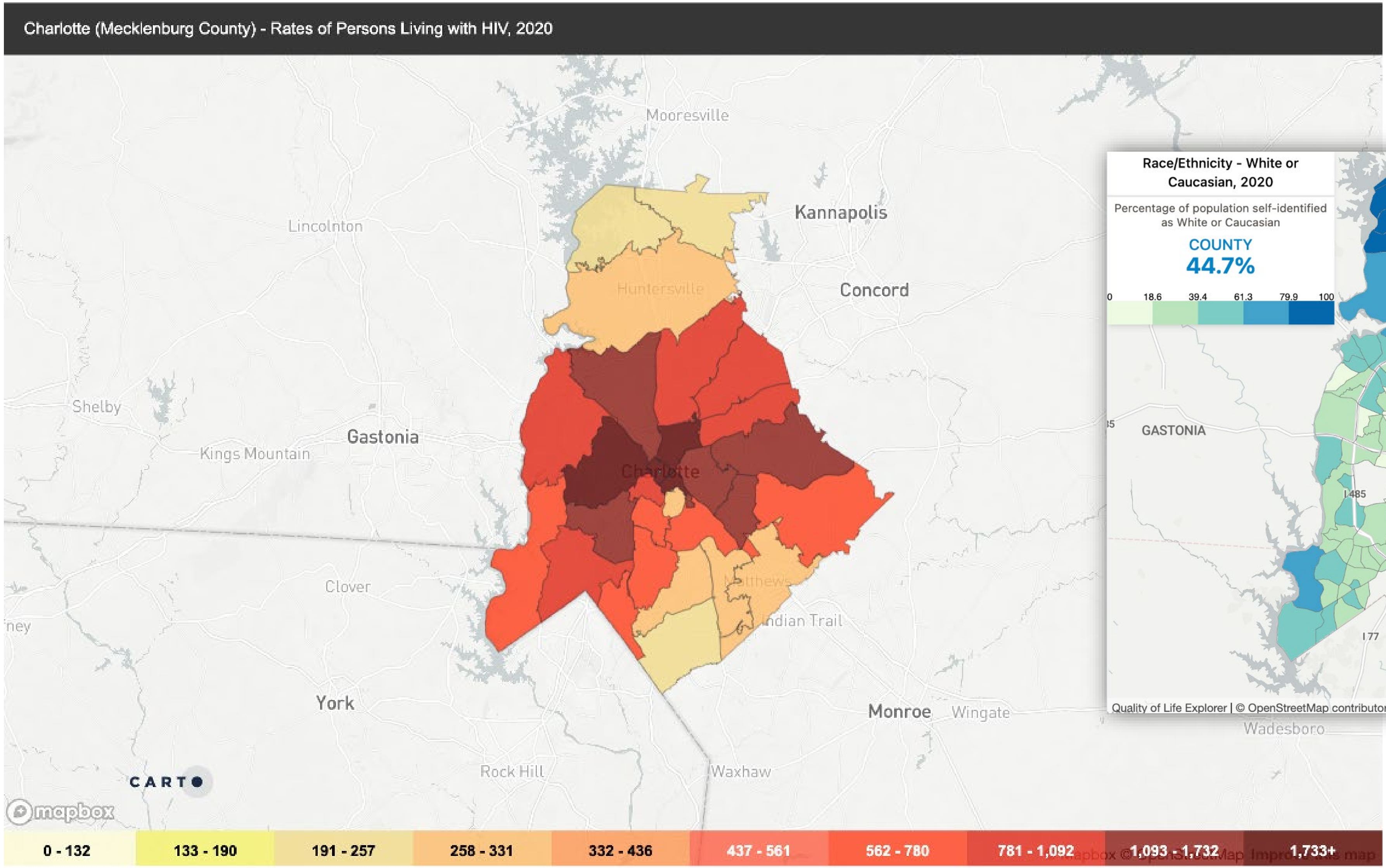
4: High risk, high benefit

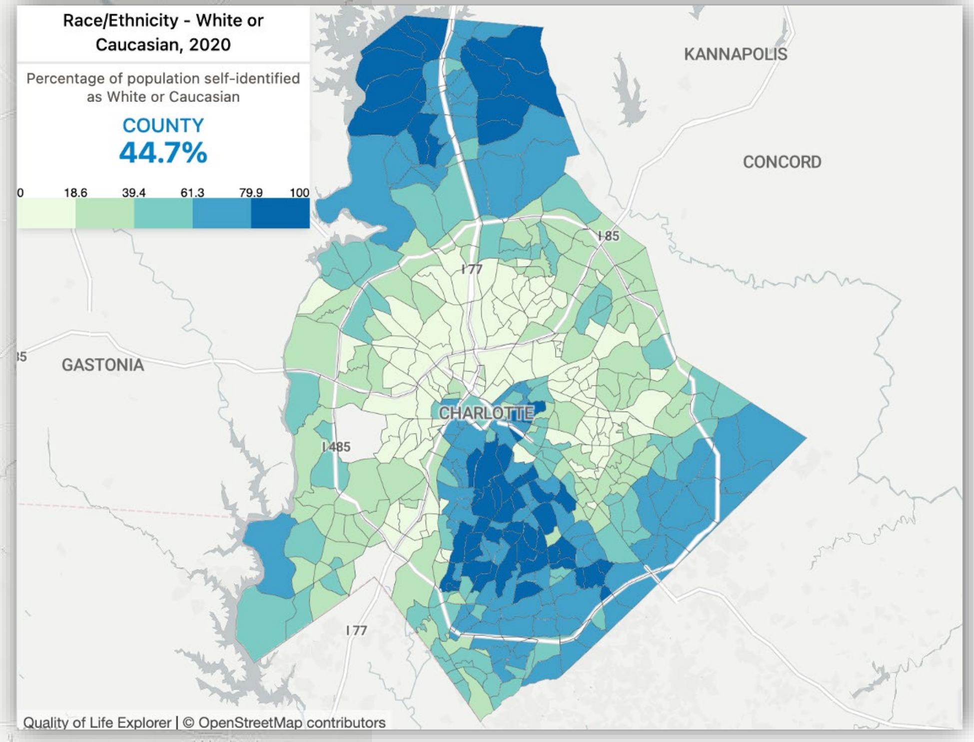
What is the risk vs. benefit?

a. HIV prevalence geocoded by zip code

b. HIV Diagnoses by neighborhood, sex, race, ethnicity

In 2020, there were 6,668 people living with HIV in Charlotte (Mecklenburg County).
In 2020, 209 people were newly diagnosed with HIV.





Sullivan PS, Woodyatt C, Koski C, Pembleton E, McGuinness P, Taussig J, Ricca A, Luisi N, Mokotoff E, Benbow N, Castel AD. A data visualization and dissemination resource to support HIV prevention and care at the local level: analysis and uses of the AIDSVu Public Data Resource. Journal of medical Internet research. 2020;22(10):e23173.

NYC OpenData

Home Data About Learn Alerts Contact Us Blog Sign In

HIV/AIDS Diagnoses by Neighborhood, Sex, and Race/Ethnicity

View Data | Visualize | Export | API | ···

Health

These data were reported to the NYC DOHMH by March 31, 2021

This dataset includes data on new diagnoses of HIV and AIDS in NYC for the calendar years 2016 through 2020. Reported cases and case rates (per 100,000 population) are stratified by United Hospital Fund (UHF) neighborhood, sex, and race/ethnicity.

Updated
March 13, 2023

Data Provided by
Department of Health and Mental Hygiene (DOHMH)

About this Dataset

Mute Dataset

Updated
March 13, 2023

Data Last Updated
March 13, 2023

Metadata Last Updated
March 13, 2023

Date Created
February 22, 2017

Views **9,601**
Downloads **2,967**

Data Provided by
Department of Health and Mental Hygiene (DOHMH)

Dataset Owner
NYC OpenData

Dataset Information

Agency	Department of Health and Mental Hygiene (DOHMH)

Update

Update Frequency	Annually
Automation	Yes
Date Made Public	4/3/2018

Attachments

DOHMHDataDictionary_Reportable_Disease_Surveillance_Data_HIV_AIDS_Diagnoses_by_Neig_Sex_Race_011118.xlsx

Topics

Category	Health
Tags	*This dataset does not have any tags*

What's in this Dataset?

Rows **8,976**
Columns **11**
Each row is a **Diagnoses of HIV/AIDS by Year, Neighborhood, Sex, and Race/Ethnicity**

NYC Open Data, HIV/AIDS Diagnoses by Neighborhood, Sex, and Race/Ethnicity, March 2023

In the Poll: What did you decide?
Where did you place a-d?

a. HIV prevalence geocoded by zip code

b. HIV Diagnoses by neighborhood, sex, race, ethnicity

Tools and Techniques

Disclosure Limitation Methods

Information limiting methods

Methods that limit or modify the amount of information available in a dataset in order to protect individual privacy.

Data perturbation methods

Methods that involve making intentional modifications to the data to prevent re-identification while maintaining the overall utility and statistical properties of the data.

Information limiting methods

Removing identifiers	Removal of all direct personal identifiers
Aggregation	Individual data entries are combined into summary statistics, such as totals, averages, or counts
Suppression	Low frequency count data and/or sensitive cells are identified and redacted
Blurring	Reduce the precision through rounding, percentages, or ranges instead of exact counts
Collapsing	Collapsing reported categories to eliminate small counts that would otherwise need protection

Data perturbation methods

Data swapping	Values of certain variables are exchanged between records.
Noise	Random noise is added to the data to obscure individual data points.

Example 1

Unsuppressed Table

Eligible for Free Meals		Eligible for Reduced-Price Meals		Not Eligible for Free or Reduced-Price Meals	
N	**%**	**N**	**%**	**N**	**%**
2	2%	0	0%	98	98%

Suppressed Table

Eligible for Free Meals		Eligible for Reduced-Price Meals		Not Eligible for Free or Reduced-Price Meals	
N	**%**	**N**	**%**	**N**	**%**
*	< 5%	0	0%	*	> 95%

masking

bottom-coding

top-coding

Example 2

Unsuppressed Table

Eligible for Free Meals		Eligible for Reduced-Price Meals		Not Eligible for Free or Reduced-Price Meals	
N	**%**	**N**	**%**	**N**	**%**
2	10%	0	0%	18	90%

Suppressed Table

Eligible for Free Meals		Eligible for Reduced-Price Meals		Not Eligible for Free or Reduced-Price Meals	
N	**%**	**N**	**%**	**N**	**%**
*	≤ 10%	0	0%	*	≥ 90%

masking

bottom-coding

top-coding

Example 3

Complementary Suppression

Student Group	Number of Students	Percent Proficient
American Indian	***	***
Asian	15	87.7%
Black	12	91.7%
Hispanic	21	81.0%
Two or More Races	13	76.9%
White	24	79.2%
Female	45	84.4%
Male	41	78.0%

Adapted from PTAC, Disclosure Avoidance and Limiting Access to PII - November 2012

Example 3

Complementary Suppression

Student Group	Number of Students	Percent Proficient
American Indian	*** (1 student)	***
Asian	15	87.7%
Black	12	91.7%
Hispanic	21	81.0%
Two or More Races	13	76.9%
White	24	79.2%
Female	45	84.4%
Male	41	78.0%

Adapted from PTAC, Disclosure Avoidance and Limiting Access to PII - November 2012

Example 3

Complementary Suppression

Student Group	Number of Students	Percent Proficient
American Indian	*** (1 student)	***
Asian	15	87.7%
Black	12	91.7%
Hispanic	21	81.0%
Two or More Races	13	76.9%
White	24	79.2%
	85	
Female	45	84.4%
Male	41	78.0%
	86	

15 + 12 + 21 + 13 + 24 = 85
45 + 41 = 86
86 – 85 = 1

Adapted from PTAC, Disclosure Avoidance and Limiting Access to PII - November 2012

Example 3

Complementary Suppression

Student Group	Number of Students	Percent Proficient
American Indian	*** (1 student)	***
Asian	15 (13 student)	87.7% = 13 ÷ 15
Black	12 (11 student)	91.7% = 11 ÷ 12
Hispanic	21 (17 student)	81.0%
Two or More Races	13 (10 student)	76.9%
White	24 (19 student)	79.2%
Female	45 (38 student)	84.4%
Male	41 (32 student)	78.0%

Adapted from PTAC, Disclosure Avoidance and Limiting Access to PII - November 2012

Example 3

Complementary Suppression

Student Group	Number of Students	Percent Proficient
American Indian	*** (1 student)	(0.0%) = 70 ÷ 70
Asian	15 (13 student)	87.7% = 13 ÷ 15
Black	12 (11 student)	91.7% = 11 ÷ 12
Hispanic	21 (17 student)	81.0%
Two or More Races	13 (10 student)	76.9%
White	24 (19 student)	79.2%
	70	
Female	45 (38 student)	84.4%
Male	41 (32 student)	78.0%
	70	

Adapted from PTAC, Disclosure Avoidance and Limiting Access to PII - November 2012

Example 3

Complementary Suppression

Student Group	Number of Students	Percent Proficient
American Indian	***	***
Asian	15	87.7%
Black	***	***
Hispanic	21	81.0%
Two or More Races	13	76.9%
White	24	79.2%
Female	45	84.4%
Male	41	78.0%

By suppressing an additional student group, reidentification of American Indian student group is prevented.

In this case, the next smallest student group, Black student group, is suppressed.

Adapted from PTAC, Disclosure Avoidance and Limiting Access to PII - November 2012

Example 4

coarsening

blurring

Record Number	Date of Birth	County	Income	Race
1	4/12/1953	Alpha	61,123	White
2	12/8/1988	Alpha	48,420	White
3	5/1/1996	Beta	30,288	Black
4	2/20/1979	Beta	52,189	White
5	1/7/1966	Beta	117,963	White
6	10/14/1972	Gamma	138,228	Black
7	7/9/1981	Gamma	103,242	White
8	3/12/1992	Gamma	45,144	White
9	8/13/1967	Gamma	62,513	White
10	12/20/1986	Delta	85,232	White

Record Number	Year of Birth	County	Income	Race
1	1953	Alpha	60,000-69,999	White
2	1988	Alpha	40,000-49,999	White
3	1996	Beta	30,000-39,999	Black
4	1979	Beta	50,000-59,999	White
5	1966	Beta	110,000-199,999	White
6	1972	Gamma	130,000-139,999	Black
7	1981	Gamma	100,000-109,999	White
8	1992	Gamma	40,000-49,999	White
9	1967	Gamma	60,000-69,999	White
10	1986	Delta	80,000-89,999	White

Federal Committee on Statistical Methodology's Data Protection Toolkit, Statistical Disclosure Limitation, Protecting Microdata, Current Methods of Protecting Microdata Files, Recoding Data

More Complex Methods

Synthetic Data	Entirely new, artificial datasets are created based on the patterns of the original data. Although synthetic data reflects the characteristics of the real data, it doesn't correspond directly to real-world individuals.
Privacy-Enhancing Technology	PETs refer to cryptographic techniques to protect privacy within data systems while allowing for greater utility of the data. PETs provide a safer and more secure way to analyze, link, and share data.
Disclosure Review Boards	This shared governance model brings together experts to review information before public release.

Common Privacy Enhancing Technologies

Secure Multiparty Computation

parties jointly compute a query on their datasets, without seeing the other's underlying data, using encryption

Secure Enclave

virtual computing workspace that enables authorized users to access sensitive data and securely conduct analysis

Differential Privacy

method for obscuring identities or attributes in the underlying record-level data by infusing results or statistics with noise

Secure Hashing

an algorithm that replaces sensitive information with a random string of characters (hash) unique to each original record in the data

MAY 15, 2025

1:00 PM ET

JOIN US for **Demystifying Privacy Enhancing Technologies** Workshop

The Disclosure Limitation Combo

Disclosure limitation methods may be used:

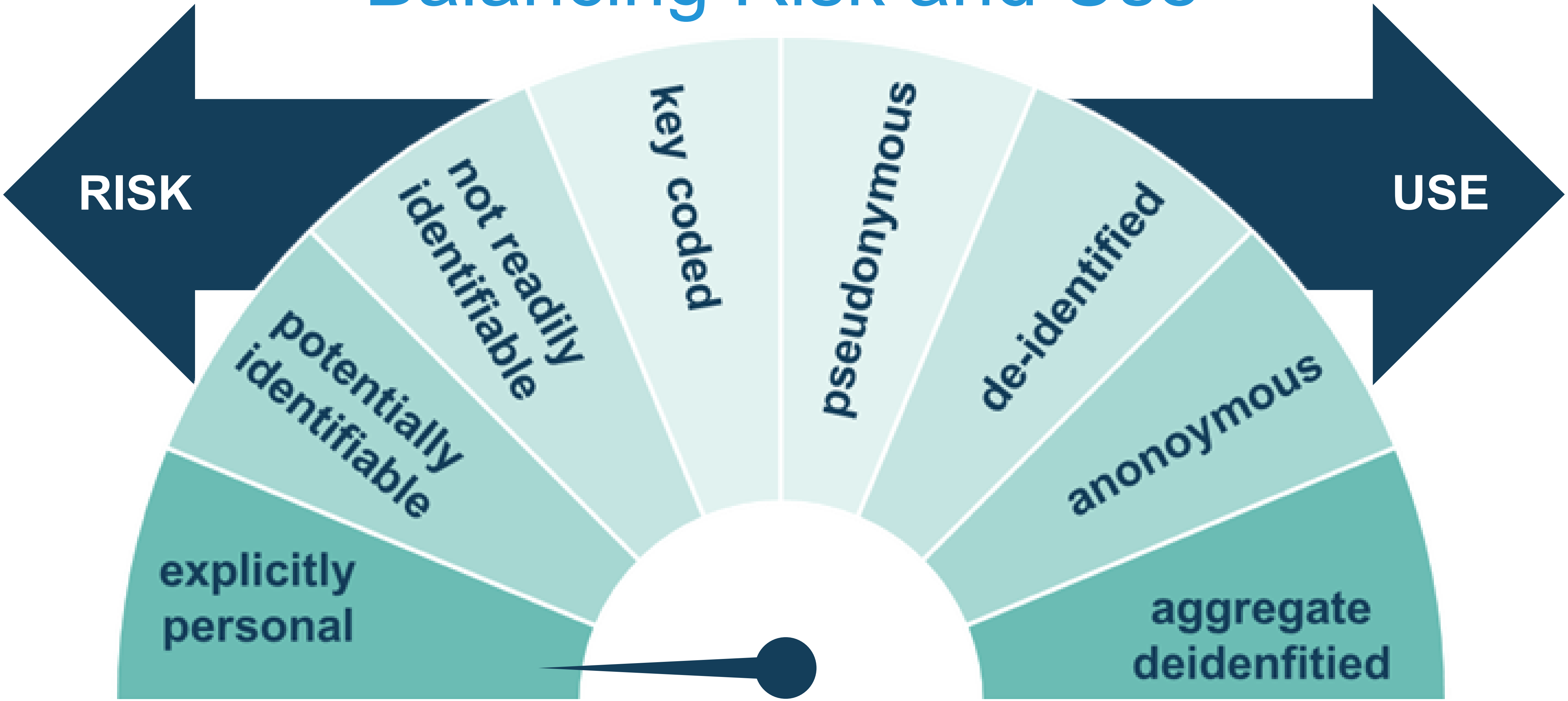
- individually or together,

AND

- as part of other administrative and technical controls.

Best Practices

Balancing Risk and Use



Adapted from [A Visual Guide to Practical De-Identification](), FPF.org

Best Practices

- **Data Minimization:** Collect and use only the data necessary for the intended analysis to reduce the risk of disclosure.
- **Anonymization and De-identification:** Apply techniques to remove or obscure personal identifiers to prevent re-identification of individuals.
- **Differential Privacy:** Employ advanced techniques like differential privacy to provide statistical insights while safeguarding individual privacy.
- **Risk Assessment:** Conduct thorough risk assessments to understand the potential for re-identification and guide the appropriate choice of disclosure limitation techniques.
- **Transparency:** Clearly communicate the methods used for disclosure avoidance to build trust and help users understand the data's limitations.

Common pitfalls and how to avoid them

Too Strict:

- Loss of Data Utility
- Misinterpretation
- Reduced Transparency
- Frustration Among Users

Too Lax:

- Privacy Breaches
- Legal and Ethical Issues
- Loss of Trust
- Exploitation of Sensitive Data

Do this:

- Policies and Procedures
- Be transparent to internal and external users
- Think through unintended consequences
- Be aware of what your data providers and partners publish

Questions?

Share your thoughts

Take a quick
[Workshop Survey]

For more trainings, visit:
https://disc.wested.org/

Thank you.

Amy Hawn Nelson
AISP
ahnelson@upenn.edu

Laia Tiderman
DISC
ltiderm@wested.org

Sean Cottrell
DISC
scottre@wested.org

A Project of WestEd Penn University of Pennsylvania